Corvinus University of Budapest

Faculty of Business Administration

BA in Business Administration and Management, year IV

Department of Operations Research and Actuarial Sciences

Predicting the Oscars with machine learning

Tudományos Diákköri Dolgozat

Supervisor: Péter Vékás

Student: Máté Váradi

March 18, 2018

Abstract

Each year before the Academy Awards are announced, several experts try to forecast the winners. I applied statistical learning models to create my own forecast for 2018 in the six main categories, and to identify the most important indicators of a win. These are primarily the results of other award ceremonies. Various kinds of Oscar related data was collected from 1960. Among the three models, the Random Forest Classifier performed best, achieving a 91.5% overall accuracy, and 100% in the 2018 season. Support Vector Machines also performed well overall. Logistic regression was useful to explain the relationship between variables and the nominated films' chances of victory. For example, the more nominations a film receives, the higher its chance is to win any award.

Keywords: prediction, Oscars, probabilities, machine learning, films, motion pictures

TABLE OF CONTENTS

TABLE OF TABLES

TABLE OF FIGURES

Introduction

Each year the Academy of Motion Picture Association (AMPAS) honours outstanding achievement in film in a wide range of categories. The first Academy Award was handed out in 1928 and the show takes place some time around late February each year ever since. The ceremony is watched live by millions around the world. For the moviegoer the Academy Award, or more commonly known, an Oscar is an indication that a film is viewed as being worthy of recognition by industry experts. For a professional, an Oscar means the industry's recognition and may result in higher salary and greater choice of roles or films in the future (Nelson, 2001).

An Academy Award may also result in additional box office revenues, mainly because it extends the time films stay in cinema, or it causes them to be rereleased (Nelson, 2001). Some researchers argue that the financial impact of an Oscar are concentrated in the nomination rather than the award itself (Deuchert et al., 2005; Simonoff, 2000), but these studies only take US box office sales into account. According to Nelson (2001), the financial rewards are significantly larger for award winners than for nominees.

Apart from the aforementioned monetary incentives for the movie industry, the general public has an interest in the outcome of the award as well. Moviegoers participate in polls and debate about who will win, similar to any popular live event. One can also easily place their bets on the Oscar on several websites. Therefore, there is plenty of reason to accurately forecast the outcome of the Oscars and to identify any trends.

To predict the outcome of the Academy Award, one must first understand how the winners are decided. The AMPAS consists of industry professionals. Membership is on an invitational basis. The Academy has more than 7000 members (Silver, 2018), among them both retired and currently active actors, cinematographers, designers, writers, etc. Membership is not limited to Americans. For instance, Ildikó Enyedi, Oscar nominated Hungarian director was also invited to join the Academy recently. The voting process goes as follows:

> "Each member can only nominate within their branch: a writer, for example, cannot submit a nomination for best sound editing. PricewaterhouseCoopers, an accountancy firm that is responsible for tallying the votes, uses a method almost

> identical to Britain's proposed 'alternative vote' system. All the first-choice ballots for each film are counted, with those above a certain threshold securing a nomination. The lowest-scoring film is then eliminated and its second-choice ballots assigned to the remaining films. The process continues until five films are over the threshold (with the exception of the best-picture category, which can have as many as ten nominees on the shortlist). If a film receives a particularly large number of nominations, so that further votes for it are in effect wasted, a trickle-down process kicks in, and subsequent ballots are redistributed to the next highest choices using a fractional weighting scheme. Once the shortlists are announced in each category, Academy members are sent a second ballot, and simply pick their favourite in each category. In this second round they are allowed to cast votes in categories outside their branch, but they are advised to avoid those where they lack expertise." (The Economist, 2015).

I created statistical learning models using Oscar related data to predict Academy Award wins in the 6 categories, that are of most public interest: Best Picture, Best Director, Best Actor in a Leading Role, Best Actress in a Leading role, Best Actor in a Supporting Role and Best Actress in a Supporting Role. The object was not to choose who should win, but to predict the behaviour of Academy voters, thus to choose who will win. For this purpose I had to select variables that have historically been useful to provide good predictions. One challenge is to find variables that differentiate Oscar winners from non-winners, and not Oscar nominees from regular films.

I collected data from 1960 onwards. The primary source for my work was a database from GitHub (Scruwys, 2017). This database was used along with a complete Oscar nominations database from Kaggle (Pandya, 2015). To fill in missing data, the Internet Movie Database (IMDB) was used. Actors' and actress' ages were queried from IMDB's names database, which contains information on practically every film industry professional (imdb.com, 2018).

I collected generic variables, that were used for prediction in each of the categories, as well as category-specific ones. I did not make a distinction in the variables I used for male and female acting categories, thus my variables are in four separate databases: one for Best Picture; one for Best Director; one for "lead acting" (Best Actor in a Leading Role and Best Actress in a Leading role) and one for "supporting acting" (Best Actor in a Supporting Role and Best Actress in a Supporting Role). When I use the word 'category' it may refer to the

six categories in the Oscars, or these four databases, since these served as the basis of my analysis.

The three models I used were: Logistic Regression; Support Vector Machine (SVM) and Random Forest. Logistic Regression was chosen because of its probabilistic nature, easy implementation and its wide usage by other researchers who predict the Oscars. Support Vector Machine was selected because they have proven to be robust and accurate in a variety of classification problems. Random Forest was chosen because it is based on decision trees, which is a very intuitive and easily interpretable method. Further advantages and disadvantages of these models will be discussed in the third chapter.

For reasons of familiarity and convenience I used Python for data cleaning, statistical analysis and model estimation. Python's Pandas and Scikit-learn packages were tremendously useful throughout the project (sklearn.org; pandas.pydata.org). All the figures in this paper are made in Python too, using Matplotlib and Seaborn (matplotlib.org; seaborn.pydata.org).

In this paper I will present my work and results of the models. First, related research will be introduced. Second, I will present the variables that I used, along with some explanatory data analysis. In the third chapter, methodology and details of the models will be discussed. Results for the 2018 season and other issues related to the outcomes of the models will be the topic of the final section of this paper.

The goal of this project is twofold. The first object is to make accurate predictions for 2018. Secondly, it is to reveal the trends and connections between Academy Award wins and other factors. How my work differs from previous research in the topic is that I tried to include the widest possible variety of variables. Instead of focusing solely on the preceding award wins, I collected numerous other data points that could be relevant. Furthermore, this is the first research paper to use Support Vector Machines and Random Forests for Oscar predictions.

1. Literature Overview

According to Scott Feinberg, awards analyst for the Hollywood News Reporter, no statistical model can substitute for what he calls "underground intelligence" about the Academy members' "subjective, whimsical preferences" (Bialik, 2013). In spite of this, several attempts have been made to predict the Oscar awards with statistical learning models. Some of these models managed to achieve considerably accurate predictions. In this section I will summarise previous research on predictive modelling of Oscar outcomes.

1.1 Logit models

D. Kaplan used a logistic regression model to predict Academy Award wins in the Best Picture category. He collected various data points: results of the DGA and Golden Globe awards, genre, and historical records of the Academy Awards. For instance he looked at how many previous nominations the director of each nominated movie had had. Kaplan's approach to using genre variables is very different to mine. He only uses five genre categories, one of which is "Epic", which is determined by Kaplan's subjective evaluation of a movie. He also included a variable which denotes the number of Oscar nominations a movie had received. His results show that the a film is more likely to win if the movie is an "epic" biography, if the film won the Golden Globe for Best Picture in the Drama cluster, and if the film won the DGA award. (Kaplan, 2006).

Iain Pardoe used McFadden's conditional logit model to predict wins in the four major categories between 1938 and 2004. His models had an overall accuracy of 69%. With more data available in the later years, prediction accuracy has improved over time. Pardoe added *age* and *age squared* variables to his model for actors and actresses and found that female winners tend to be younger, than male winners. Between 1928 and 2006, the median age of Best Actress Oscar winners was 33 years, whereas for Best Actor it was 42 years. Over time, both actor and actress nominee ages have increased. Winners tend to be slightly older than losing nominees in the recent years. Incorporating age and age squared variables failed to improve prediction of winners, however. Pardoe also found that previous Oscar nominations benefit actors and directors, whereas a previous win reduces the chance of another win for both actors and actresses. (Pardoe, 2005; Pardoe-Simonton, 2008).

1.2 Other forecasting methods

David Rothschild et al. (2014) compared four different methods to forecast winners in all 24 categories. Their goal was not only to predict the winner for a given year, but to create an *accurate, timely* and *cost effective* forecast. One such method was aggregating experts' forecasts. The least accurate method among the four was non-representative polling, where the researchers considered polled fractions of the individual nominees as probabilities of their winning. The explanation for lower accuracy with this type of forecasting might be that poll responses often reflect what respondents want as an outcome, rather than what they realistically think the outcome will be.

However, when people are incentivised to give their best guess for an outcome (e.g. by a bet), aggregating their responses can yield extremely accurate predictions. This is the idea of the wisdom of crowds. According James Surowiecki's book with the same title, large groups' aggregated answers to questions are often better than any single expert's answer (Surowiecki, 2005). Rothschild's most accurate forecasting method is based on the same idea: "Prediction markets are markets where users can buy and sell contracts on upcoming events; the price of the security is highly suggestive of the probability of the outcome. For example there was security for Daniel Day Lewis to win Best Actor that would be worth $1 if he won and $0 if he lost. Since he was extremely likely to win, people were willing to pay nearly $1 for the security, demonstrating their subjective probability was approaching 100%" (Rothschild et al., 2014, p. 13). Aggregating and transforming raw prices from three different prediction markets yielded the most accurate predictions in their study, which is in unison with Arrow's finding about prediction markets, in that such markets may produce forecasts of event outcomes with a lower prediction error than conventional forecasting methods (Arrow, 2008).

Rothschild calls his next method, "fundamentals-based", which is essentially the kind of predictive modelling that I did. Fundamental data is data that researchers do not need to collect; they exist for other reasons. Rothschild used a logit model with L1-penalized log likelihood, since many of the variables were irrelevant. He constructed 24 models in 6 time points with different data availability (e.g. when the first model is run, no other awards winners are announced yet). Altogether that is 144 models, with domain specific variables

for each category. This makes fundametals-based forecast quite costly compared to the other methods. Rothschild found, that data earlier than 1995 only made predictions less accurate. They used a *release date* variable, which denoted the days between the Oscar's ceremony and the US release of the film. Rothschild's results show that a film is more likely to get a nomination if it opens later in the season, however, once nominated, early released movies are a little more likely to win. They also collected ratings data (critical and popular), which in most cases had no predictive power. (Rothschild et al., 2014).

1.3 Non-academic modellers

Apart from the academic researchers, there are several online bloggers and journalists who also use statistical learning techniques to create their own predictions. Unfortunately they often only focus on the predictions and tend not to give details about the methodologies they use. For these reasons, I do not attempt to introduce all of them, and what each of them exactly does. Instead I will only concentrate on the modellers who use similar techniques to mine.

One well known platform for predictions on politics, sports and entertainment is *PredictWise*. *PredictWise* was founded by David Rothschild, whose methods were already discussed in the last subchapter. PredictWise makes predictions in all 24 of the categories each year, usually getting about 80% of them right. (Rothschild, 2018).

Licensed under The New York Times, *FiveThirtyEight* is a blog that focuses on opinion poll analysis, politics and economics. Its creator and editor Nate Silver attempts to use his experience from forecasting elections to predict Academy Award wins. His model relies solely on the other awards that are given out before the Oscars. He assigns a score for each film in a given category, and chooses the winner based on this score. The awards that make up the score are weighted based on the square of their historical success rate, which is doubled for awards whose voting memberships significantly overlap with the Academy. (Silver, 2013).

Lastly, Ben Zauzmer is a mathematician from Harvard working on predictions for the Hollywood Reporter. He does not give details about his mathematical formula, but it is based on movies' Metacritic and Rotten Tomatoes scores, as well as their performance in other award shows (Zauzmer, 2018).

2. Data

There is a considerable degree of uncertainty in the movie industry. For instance, a director's previous success is no guarantee of the popularity, financial success or critical acclaim of the same director's new movie. In this extremely uncertain industry, however, the Oscars can still be relatively accurately predicted. This is in part due to the fact, that the Oscars are preceded by a number of other awards that are very often awarded to the same movies, sometimes even by voters coming from the same circle of industry professionals.

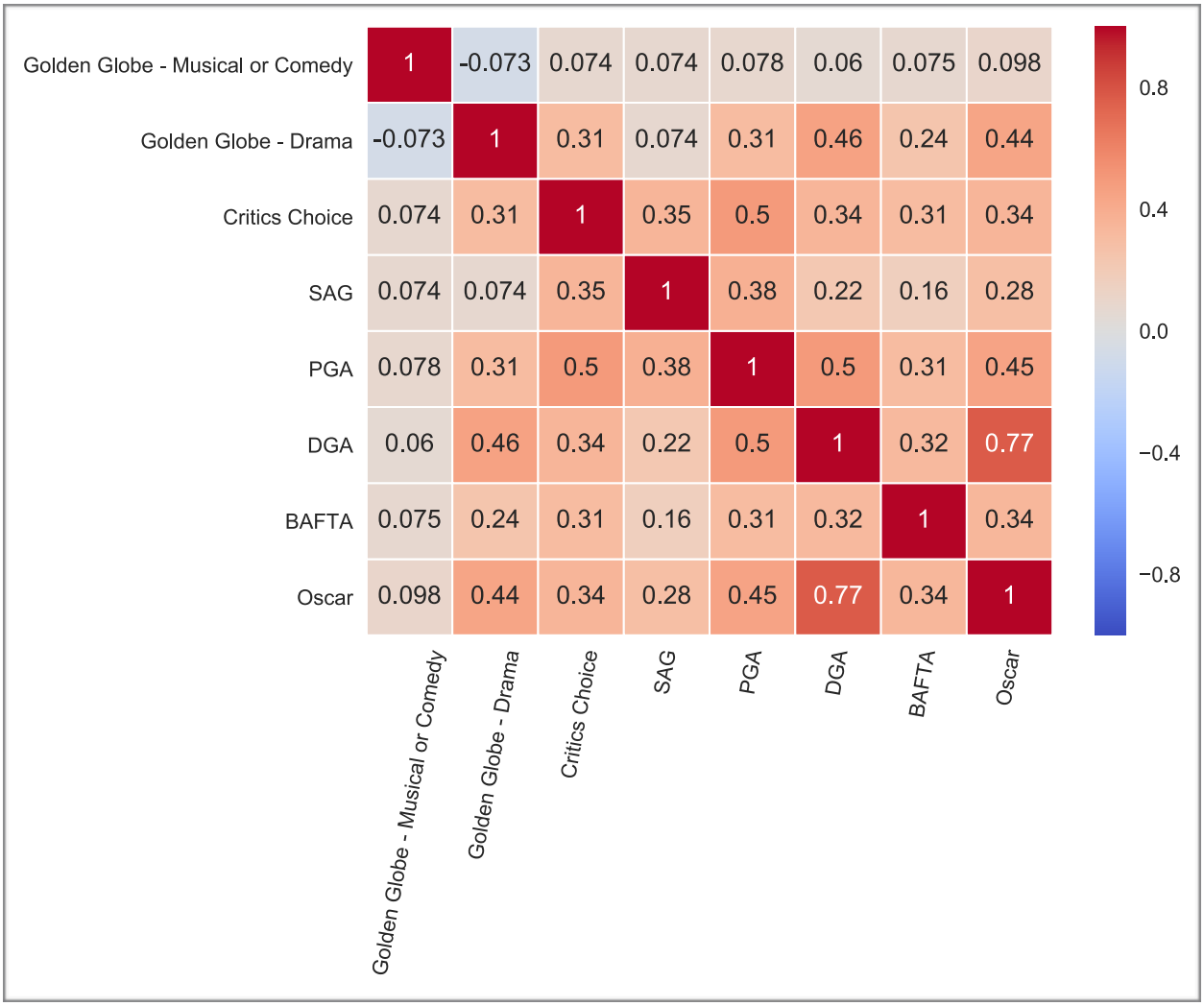

Figure 1. Correlation between awards

I collected data from all the award ceremonies displayed on Figure 1, because according to Silver (2013) these are the awards that show the most correlation with the Academy's Best Picture award. There are two types of award ceremonies. The first type is

where awards are handed out in a wide range of categories, including the six that I'm analyzing. Award shows belonging to this category are the Golden Globes, the BAFTA and the Critics Choice Movie Award. The Golden Globes distinguishes movies in the 'Drama' cluster and the 'Musical and Comedy' cluster, meaning that each year two films win an award for best movie, and two actors and two actresses win an award for 'Best Acting in a Leading Role'. Both clusters' results are used, but the Drama cluster's results tend to be more predictive for the Oscars. In the case of the first type of ceremonies, I'm using variables for award wins and nominations in the associated categories. For example to predict the Best Actor in a Supporting Role Oscar in 2018, I'm looking at the 2018 list of nominees and the final winner for the Best Actor in a Supporting Role BAFTA.

The other type of ceremonies are the Guilds, which are domain specific: the Directors Guild of America (DGA) is given to the best director of the year; the Screen Actors Guild Awards (SAG) are given to actors and actresses. The Producers Guild Award (PGA) is given to to the producers of a film rather than the film itself, they nevertheless serve as useful Best Picture precursors. Some of the Guilds can be used for prediction in multiple categories. For example, I used DGA for prediction in both Best Director and Best Picture categories. In fact, in both of these categories, a DGA win is the strongest indicator that a film will win the Oscar. As one can see from Figure I, this award shows strong correlation with the Best Picture Oscar. Screen Actors Guild has an award for "Outstanding Performance by a Cast in a Motion Picture" which can be used for prediction for all of the acting categories and for the Best Picture category as well.

In the case of both the Guilds and the first type of ceremonies, I have separate variables for wins and nominations. Sometimes the nominations themselves can be predictive. For example in the Best Director category, BAFTA and Golden Globe wins as well as nominations are useful predictors.

PGA exists since 1990, SAG since 1995 and Critics' Choice since 1996. Therefore these variables needed to be coded as a dummy, where the 3 categories are: *award non-existent; award exists, but not won; award won.* For this reason there are two categorical variables in the case of these awards, coded consistently as: *award non-existent:* [0,0]*; award*

exists, but not won: [1,0]*; award won:* [1,0]. The rest of the variables that represent award wins or nominations can simply be represented with a single dummy variable.

Most bloggers and journalists mainly look at some combination of the above awards to make their predictions. But there are more aspects that are worth analysing. I collected genre data from the Internet Movie Database (IMDB). IMDB assigns a maximum of three genres to each film. For example in the case of the movie *Interstellar* they are: Adventure, Drama and Sci-fi. I created a variable for all 18 genres to indicate a movie's exact genre. Figure 2 shows the distribution of winner and non-winner movies belonging to each genre. The Drama genre is not shown on the figure, as the vast majority of Oscar nominated movies are dramas. The Film-Noir variable had to be dropped, because no film belongs to that genre. Films classified as Biography, History, Music, Western or War get more wins in the examined six categories than other genres. Horrors and Mystery movies, on the other hand very rarely get an Oscar. Genre variables only proved to be predictive in case of the Best Director category. Directing a Western movie increases the chance of getting an Oscar, whereas a Mystery classification decreases it.

Other generic variables include MPAA-ratings and release date. MPAA is a film rating system used in the United States to rate a film's suitability for certain audiences based on its content. Today the existing categories are G (General Audiences), PG (Parental Guidance Suggested), PG-13 (Parents Strongly Cautioned), R (Restricted) and NC-17 (Adults Only). There were no NC-17 movies in my datasets, but some of the older movies followed a different rating system. These ratings had to be dropped or converted (in case the used scale was convertable).

Release date clearly is an important factor in the film industry. In my models, it is broken down to the four quarters of the year. It is apparent that most Oscar nominated movies go into cinemas in either Q1 or Q4, meaning the first quarter of the ceremony year or the last quarter of the previous year: between 1960 and 2018 among the films nominated in the six categories that I analyzed, 30% was released in Q1 and 42% in Q4. The Academy Awards ceremonies are held typically around late February, and movies that have a good chance of being nominated for the Oscar usually postpone their release until December or January in order to be able to benefit financially from an Oscar nomination or win (Nelson,

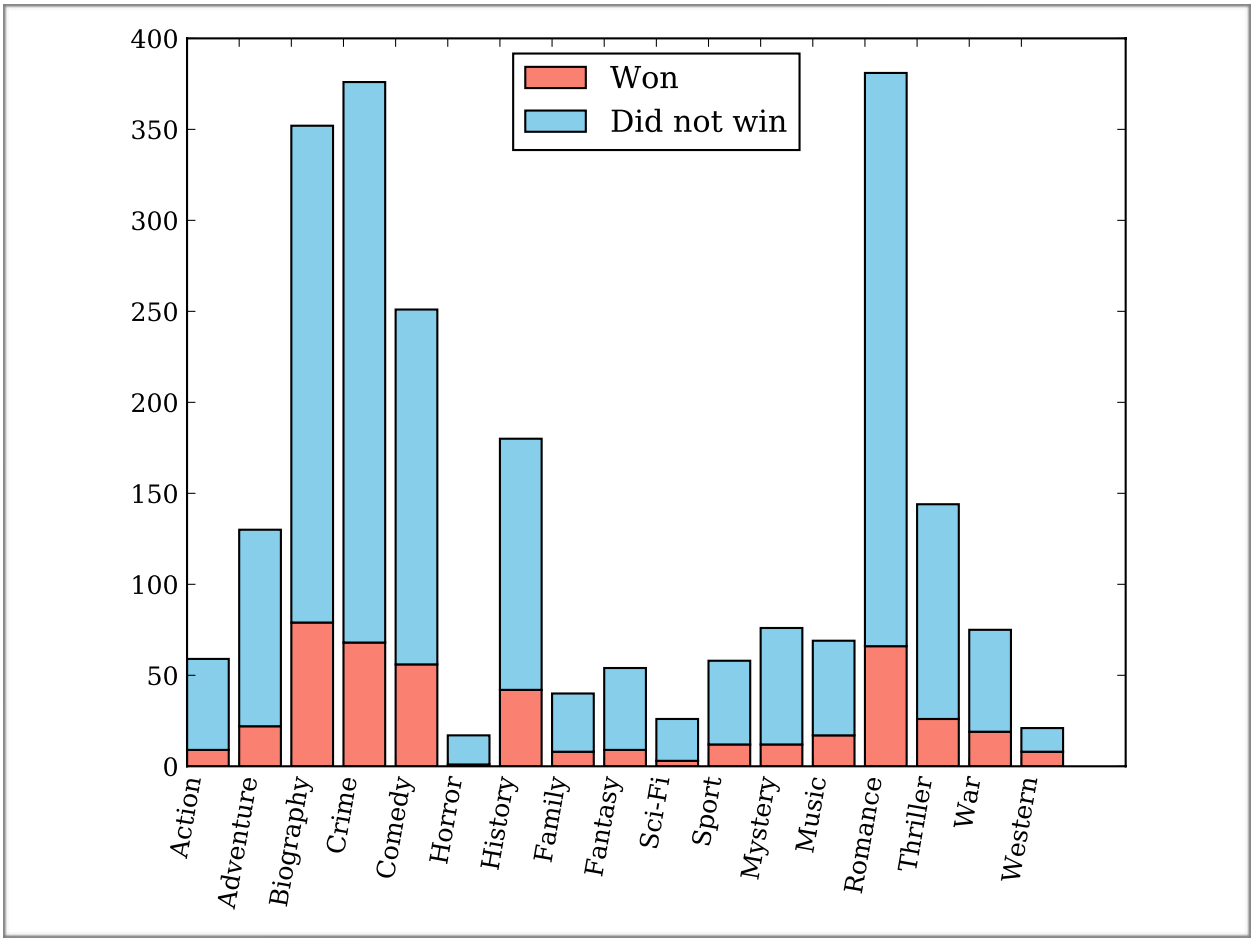

Figure 2. Genre distribution of Oscar nominated films between 1960 and 2018

2001). Furthermore, according to Simonton there is a correlation of 0.12 (p=0.001) between critical acclaim and release date. The later in the year a film is released the higher is its critical evaluation (Simonton, 2017, p. 14). Including MPAA and release date variables did not improve prediction accuracy however. Rothschild's release date variable proved to be predictive, ergo it is possible that including release date in the form that I did (with four dummies for the four quarters) did not carry enough information to manifest the real effect of the release date.

Most researchers agree that there is little or no association between budget and most important movie awards (Simonton, 2005 as cited in Pardoe, 2008). Success at the box office is hard to measure accurately and according Rothschild et al. (2014) it has little predictive power. For these reasons I did not include budget or box office variables.

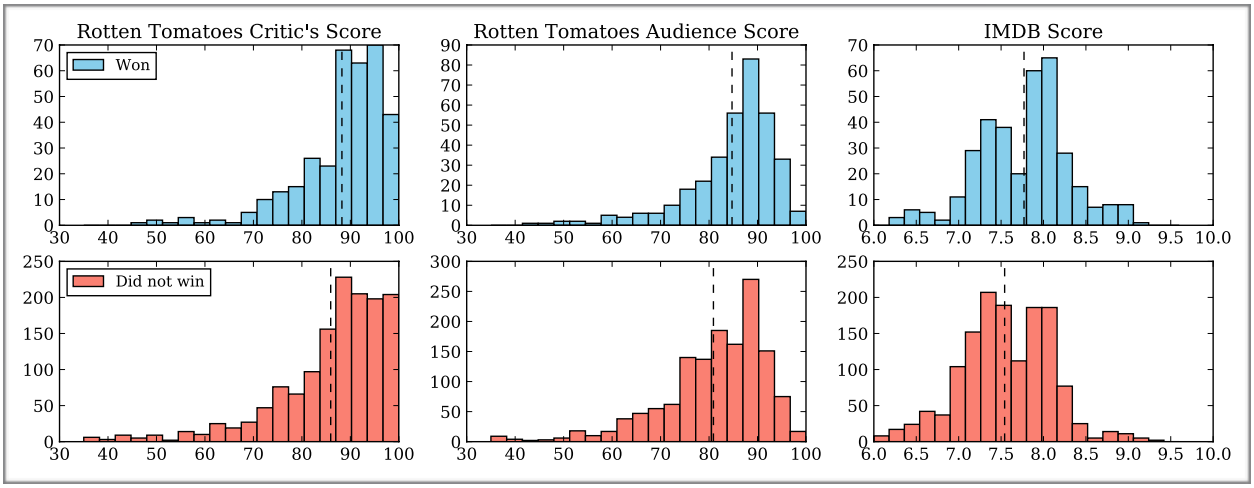

Figure 3. Distribution of Rotten Tomatoes and IMDB scores

The effect of critical and popular ratings is not so clear, however. The best way to measure critical acclaim is through Metacritic score, which is the weighted average of many reviews coming from reputed critics. However, Metacritic was created in 1999, and for several movies before that year, it simply doesn't exist. I used Rotten Tomatoes' Critics Score as an alternative. To measure popular acclaim I used IMDB score and Rotten Tomatoes' Audience Score.

Figure 3 shows that on average winners tend to have a higher score on all three of these scales. Rotten Tomatoes' Audience Score appeared to be a useful predictor, slightly increasing the chance of winning in all categories but one. In supporting acting categories both the Audience Score and the Critics' Score have a significant effect: the earlier decreases and the latter increases the chance of winnng.

Some of the most important trends can be read out from the history of the Academy Awards. Only four movies have won the Best Picture Oscar without also receiving a Best Director nomination (*Wings* in 1928, *Grand Hotel* in 1932, *Driving Miss Daisy* in 1989 and *Argo* in 2013). Conversely, only two directors have won a Best Director Oscar for a movie that was not nominated for Best Picture (Lewis Milestone for *Two Arabian Nights* in 1928 and Frank Lloyd for *The Divine Lady* in 1929). 63% of the acting awards between 1960 and 2018 have gone to actors and actresses in a movie that was also nominated for Best Picture.

A variable that was found to be useful for prediction in all categories is the number of total nominations. The median number of Oscar nominations for films that win the Best

Director award is 10, whereas for losers it is only 6. Similar patterns can be observed in the other categories as well: 9 versus 6 for Best Picture and 6 versus 4 for all acting awards combined.

I also included category-specific variables. For directors and actors I analyzed their previous Oscar nominations and wins. Directors have an increased chance of winning, the more nominations they have received in the past, and a decreased chance of winning the more times they have won in previous years. Concerning actors and actresses, both previous nominations and wins seem to decrease their chances of winning.[1] Between 1960 and 2018 22% of acting nominees have won the Oscar with no previous wins, whereas 14% of acting nominees won with one or more previous wins. The same pattern can be observed for previous nominations, although the difference is not as substantial.

Inspired by Pardoe (2008), I collected the age of actors and actresses too. Pardoe's *age* and *age squared* variables did not improve his models, therefore I took a different approach. I created seven age segments for those below 25; those above 75; and five equally sized segments between them. The 35-45 segment has the most amount of winners, but it is the nominees above 75 who have the greatest probability of winning: 23,7% of them win an acting Oscar, whereas only 11% of the actors and actresses younger than 25. Incorporating the age segment variables along with the other category-specific variables did not improve prediction accuracy. [2]

[1] This partly contradicts Pardoe's findings. He found that for Lead Actors, a previous nomination increases the chance of their winning (Pardoe, 2008).

[2] This and all the similar conclusions in Chapter 2 are based on the coefficients and p-values of the logit model. In the other two models, the effect of the features is not as transparent.

3. Model Specification

I created three separate models yielding their own predictions in all 6 categories. All three models were trained on 70% of all the observations, and the rest of the observations were used for testing. The training samples were selected randomly, regardless of the year. I will discuss these models in more detail, but to be able to compare them, first I will introduce the metrics that I used to measure prediction accuracy.

3.1 Accuracy measures

Predicting Oscar winners is essentially a binary classification problem, meaning that I have to identify to which of set of categories (winner or loser) a new observation (a newly nominated film) belongs. However it is a fairly special problem, when it comes to model evaluation. For this reason I used two metrics to evaluate model performance.

3.1.1 Sensitivity

The number of nominees varies each year in each category, but usually there are up to 10 nominees for Best Picture, and 5 nominees in all the other categories that I'm analysing. In all three models, I assign the winner prediction to the film with the highest estimated probability of winning per year, and the rest are classified as to-be-losers. Naturally there is only one winner, therefore failing to get one prediction right in one category in a given year yields two incorrect classifications. Yet in my opinion, in this problem - since winner detection is my goal - the focus should be on how often the model gets the winner right. Hence, the first metric I used to evaluate model performance is simply the true positive rate, or sensitivity:

$$TPR = \frac{TP}{TP + FN}. \tag{1}$$

The values in Table 1 were calculated after the models were fit to the entire dataset and not only to the test set, so that they can be interpreted intuitively as the percentage of years [3] where a model yielded correct predictions.

[3] The number of years *n* is equal in all categories (*n=58)*, but there are two separate Award categories (for males and females) merged together in both the Lead Acting and Supporting acting categories.

Table I. *True positive rates per category for each model*

Sensitivity	Logistic Regression	Random Forest Classifier	Support Vector Machine
Best Picture	79%	93%	85%
Best Director	95%	98%	91%
Lead acting categories	71%	92%	84%
Supporting acting categories	53%	87%	86%

n=58

3.1.1 Receiver Operating Characteristic

The receiver operating characteristic, i.e. ROC curve, is a graphical plot that illustrates the performance of a binary classifier. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The overall performance of a classifier, summarised over all possible thresholds, is given by the area under the curve (AUC). An ideal ROC curve will hug the top left corner, so the larger the AUC the better the classifier. ROC curves are useful for comparing different classifiers, since they take into account all possible thresholds. (Tibshirani et al., 2013).

ROC curves were fit on the test sets. From Table I and Figure 4 we can derive the most important conclusions about the overall predictability of the categories and the performance of the models. In most categories the random forest classifier performs best: for example, it makes only one misclassification between 1960 and 2018 in the Best Director category. This category appears to be the most predictable, while the outcomes of the supporting acting categories tend to be slightly more difficult to guess.

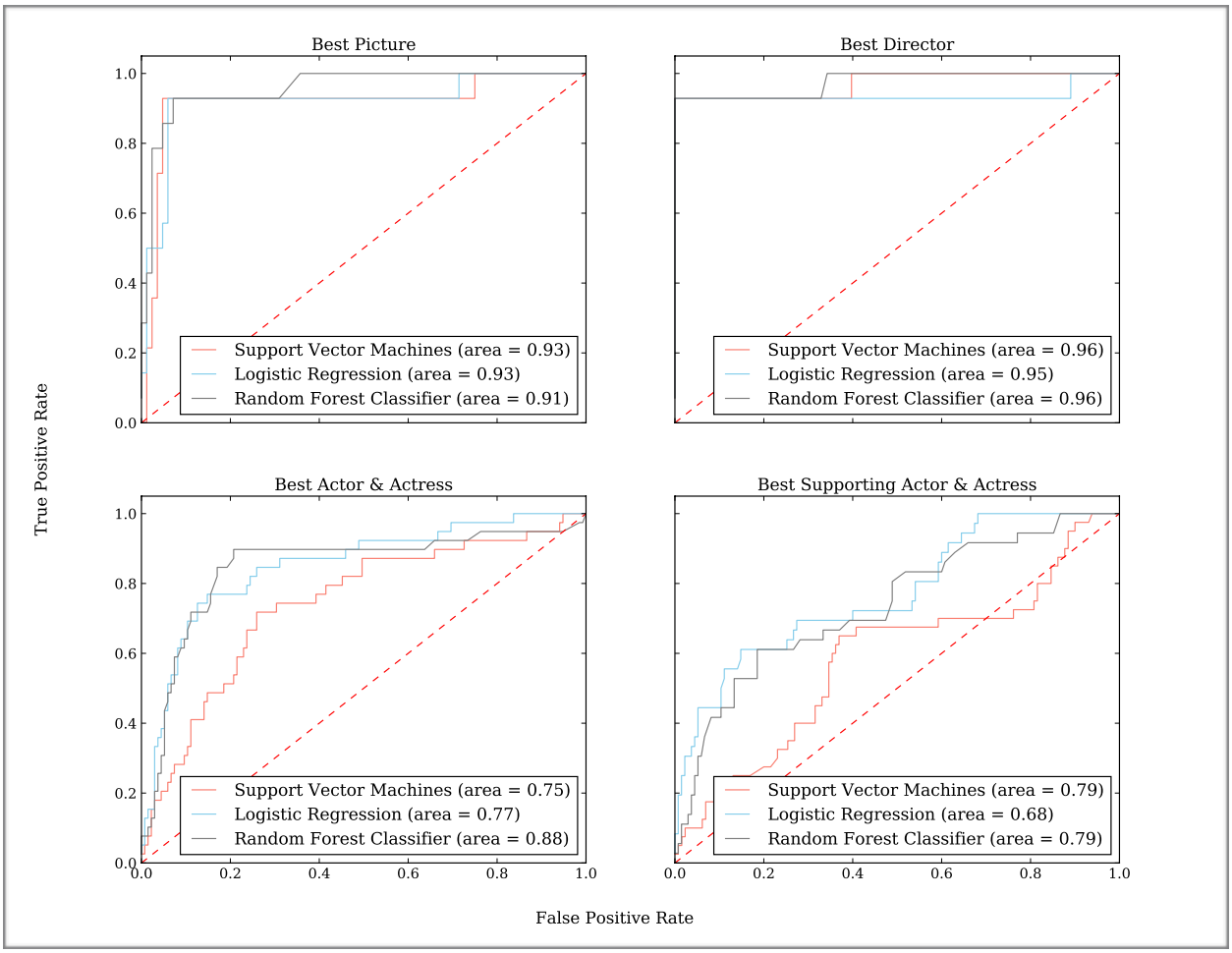

Figure 4. ROC curves comparing the performance of the three models

3.2 Logistic regression

Logistic regression is a linear model for classification. It estimates the function:

$$p(y = 1 \mid \mathrm{x}) = \frac{e^{\beta_0 + \beta^T \mathrm{x}}}{1 + e^{\beta_0 + \beta^T \mathrm{x}}}, \tag{2}$$

where $\mathrm{x} = (x_1, x_2, \ldots, x_p)$ are p predictors and $p(y = 1 \mid \mathrm{x})$ is the conditional probability of the event $y = 1$ given the observed predictors, which in my case is an Oscar win for a given movie. Python's Scikit-learn package uses L2-regularization by default (Pedregosa et al., 2011). This means that instead of using maximum likelihood estimation, where the objective is

$$\min_{\beta} \sum_{i=1}^{n} - \log\, p(y_i \mid \mathrm{x}_i; \beta), \tag{3}$$

it estimates coefficients by solving

$$\min_{\beta} \sum_{i=1}^{n} - \log\ p(y_i | x_i; \beta) + \frac{\lambda}{2} \left\| \beta \right\|^2, \tag{4}$$

where $\lambda \geq 0$ is the regularization strength, a tuning parameter (Raschka, 2016). C is the the inverse of λ. Smaller C values specify stronger regularization. Scikit-Learn's *LogisticRegressionCV* finds the best C value using cross validation. (Tibshirani et al., 2013).

I originally had many explanatory variables (the most is p=60 for the Lead acting categories). According to Sachan (2015), logistic regression doesn't perform well when the feature space is large. Therefore I implemented feature selection with Sklearn's *SelectKBest*. This method is based on F-test estimates of the degree of linear dependency between two random variables (Pedregosa et al., 2011). I selected and used the 10 best predictors in all four categories, except for the Best Director category. In this category, using only 10 explanatory variables appeared to have a negative impact on prediction accuracy, therefore I used the best 15 predictors instead. The resulting variables and their coefficients can be seen in Table 2 and Table 3.

Table 2. *Applied features of the logit model for the leading and supporting acting categories*

Leading acting categories			Supporting acting categories		
Variable name*	Coefficient	p-value	Variable name	Coefficient	p-value
rt_audience_score	0,0231	0,0001	rt_audience_score	-0,0030	0,0653
total_oscar_noms	0,0469	0,0000	rt_critic_score	0,0231	0,0422
best_film_nom [Yes]	0,5551	0,0004	total_oscar_noms	0,0497	0,0025
SAG_win_1 [Yes]	-1,2635	0,0000	SAG_win_1 [Yes]	-0,7086	0,0013
SAG_win_2 [Yes]	2,3804	0,0000	SAG_win_2 [Yes]	0,9888	0,0000
BAFTA_nom [Yes]	0,0831	0,0041	BAFTA_nom [Yes]	0,3434	0,0233
BAFTA_win [Yes]	1,3318	0,0000	BAFTA_win [Yes]	0,3147	0,0000
critics_choice_win_1 [Yes]	-0,2671	0,0072	critics_choice_win_1 [Yes]	-0,3880	0,0069
critics_choice_win_2 [Yes]	0,1336	0,0000	critics_choice_win_2 [Yes]	0,3179	0,0001
GG_drama_lead_win [Yes]	1,8798	0,0000	GG_supporting_win [Yes]	1,8449	0,0000

Table 3. *Applied features of the logit model for the Best Director and Best Picture categories*

Best Director			Best Picture		
Variable name*	Coefficient	p-value	Variable name	Coefficient	p-value
rt_audience_score	0,0383	0,0074	rt_audience_score	0,0175	0,0098
total_oscar_noms	0,2132	0,0000	total_oscar_noms	0,2438	0,0000
best_film_nom [Yes]	0,9501	0,0687	SAG_nom_1 [Yes]	-5,0466	0,0020
DGA_nom [Yes]	0,9842	0,0618	PGA_win_1 [Yes]	-0,1940	0,0017
BAFTA_nom [Yes]	0,8979	0,0102	PGA_win_2 [Yes]	1,0347	0,0000
critics_choice_nom_1 [Yes]	-1,1709	0,0183	SAG_win_2 [Yes]	1,9827	0,0008
critics_choice_nom_2 [Yes]	0,7682	0,0415	DGA_win [Yes]	4,0063	0,0000
DGA_win [Yes]	3,7821	0,0000	BAFTA_win [Yes]	0,4330	0,0000
BAFTA_win [Yes]	0,3950	0,0001	critics_choice_win_2 [Yes]	-0,2380	0,0000
critics_choice_win_1 [Yes]	-0,8222	0,0187	GG_drama_win [Yes]	1,1530	0,0000
critics_choice_win_2 [Yes]	0,4195	0,0000			
gg_win [Yes]	0,1260	0,0000			
gg_nom [Yes]	0,0142	0,0156			
mystery [Yes]	-0,4939	0,0716			
western [Yes]	0,7050	0,1606			

*A full list of the explanatory variables and their description can be seen in Appendix A

The modelling process consists of fitting the logit model on the training sample first in each category. Second, the model is fit to each nominated film in each year. Third, predictions are made by selecting the film with the highest probability of winning in every category for each year.

3.2 Random Forest classifier

Random forests are built from decision trees. A decision tree consists of nodes. The node on the top is the "root" of the tree, which has no incoming edge. The other nodes have exactly one. A node that has outgoing edges is called an internal node, and the ones that do

not are called leaves. At each of the internal nodes the instance space is split into two (or more) sub-spaces according to one of the input attribute's value (Rokach, 2005). In the classification setting the predicted class is the most common class in the terminal nodes. One of the frequently used criterions for making the splits is the Gini index, defined by

$$G_m = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) \tag{5}$$

where $\hat{p}_{mk}$ represents the proportion of training observations in the mth region that are from the kth class. The Gini index measures node purity: a small value indicates that a node contains predominantly observations from a single class. (Tibshirani et al., 2013).

The main advantages of using classification trees include easy interpretation, the ability to display them graphically and the fact that trees can handle qualitative predictors without the need to create dummy variables. Since I needed to create dummy variables for the other models nevertheless, I did not take advantage of this last mentioned attribute. A further point is that trees are believed to more closely mirror human decision making (Tibshirani, 2013).

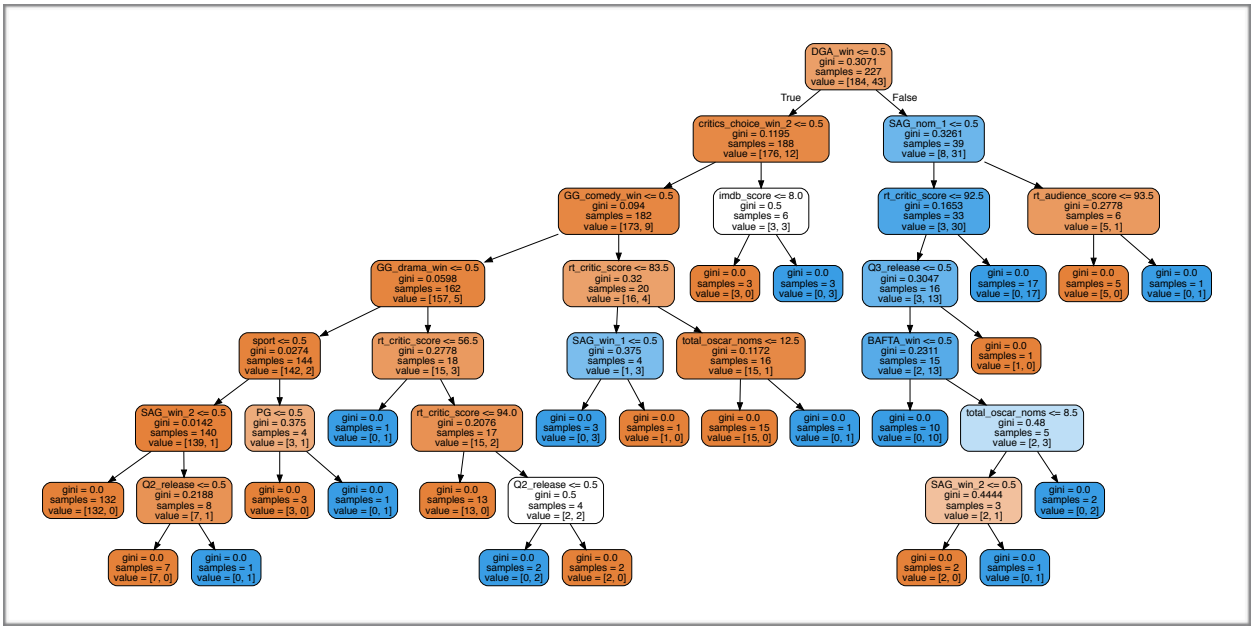

Figure 5. Decision tree for the Best Picture category

On Figure 5 we can observe a single decision tree fit onto the Best Picture training set. The training set consists of 227 films. The most important predictors can be found on the top of the tree and feature importance is decreasing as we move down on the tree. Blue nodes indicate that the most common class in the node is the 'winner' class, whereas orange nodes express a 'loser' majority. Nodes with lower purity are coloured with a lighter colour: white nodes signal a Gini index of 0.5.

The above tree is included for visualization purposes. It is not the model I used to yield predictions. I used a random forest, which is an *ensemble* of trees. Random forests were first introduced by L. Breiman. In his definition

> "a random forest is a classifier consisting of a collection of tree-structured classifiers $\{h(x, \Theta_k), k = 1,...\}$ where the $\{\Theta_k\}$ are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x " (Breiman, 2001).

Therefore, a random forest classifier is created by building a number of decision trees on bootstrapped training samples, and each time a split is made only a random sample of m predictors is chosen from the all the available p predictors. Python's *sklearn.ensemble* package considers $m = \sqrt{p}$ predictors at each split. The rationale behind considering only a subset of predictors is that this way we avoid that each tree uses the same strong predictors. Without this step the trees would be very similar and therefore their predictions would be highly correlated. Selecting from only $\sqrt{p}$ predictors *decorrelates* the trees, which consequently leads to more reliable results. (Tibshirani et al., 2013). I used n=250 estimators (trees) in my random forest.

To summarize, in this model each tree votes whether the input films are winners or losers, and then these votes are aggregated. The most common vote is the final prediction for a film. The random forest classifier's success in predicting Oscar wins might be due to it's process being similar to how the winners are decided in real life: by aggregating votes. Another advantage of using random forests, which might be beneficial in this case is that they take into account variable interactions (Sachan, 2015).

As mentioned earlier, the variable on top of the trees tend to be more important for prediction. The feature importance for variable x_m can be quantified by adding up the

weighted impurity decreases for all nodes where x_m is used, averaged over all n trees in the forest. When using the Gini index as impurity function, this measure is known as the Gini importance or Mean Decrease Gini. (Breiman, 2001).

Acquiring feature importances can be performed in Python's Scikit-learn package. These are displayed on Figure 6 for the Best Director category. The distribution of feature importances is skewed: relatively few variables actually matter and the rest are not so important. It is interesting to compare the most important variables here with the logit model's best predictors. A few discrepancies can be found, for example the `imdb_score` and `rt_critic_score` variables did not seem important in the logit model for predicting the outcome of the Best Director category, but they are in the random forest classifier.

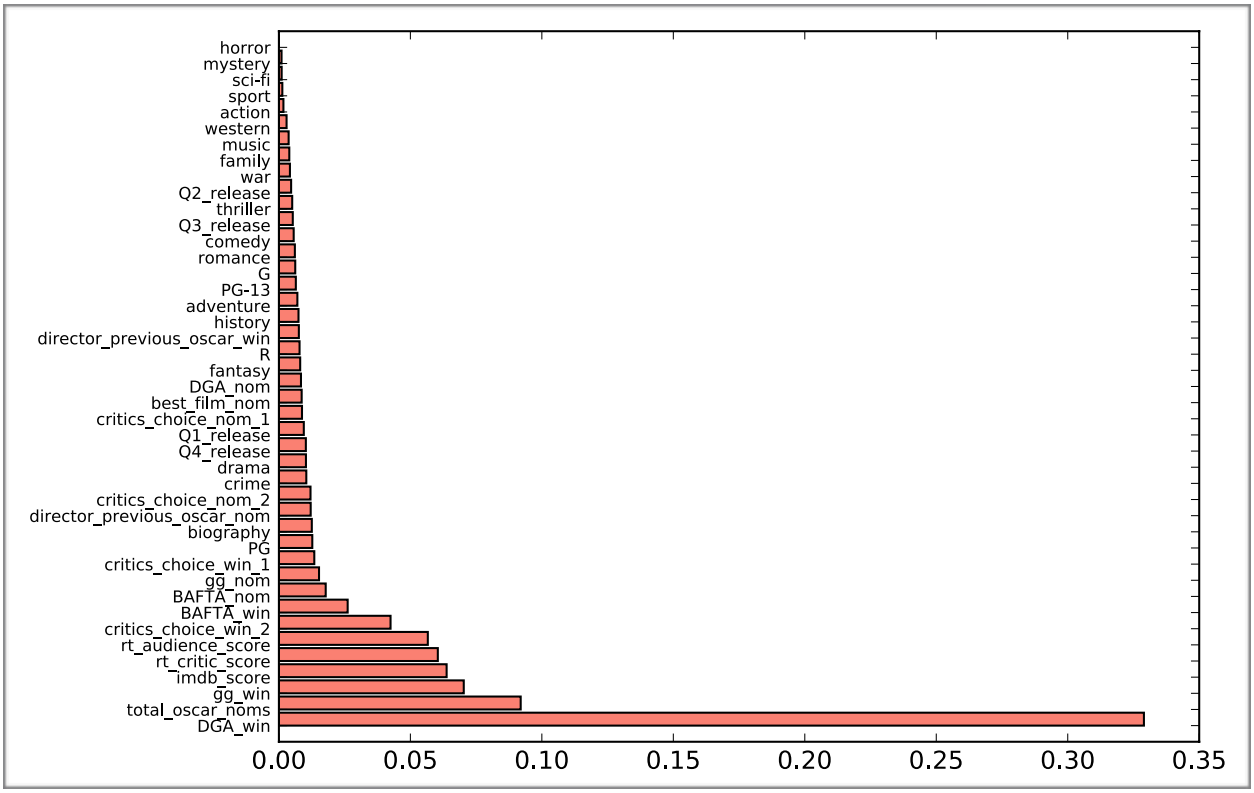

Figure 6. The Random Forest Classifier's feature importances for the Best Director category

The Best Director category was chosen at random to visualize the distribution of feature importances. For each category the 20 most important variables and their Mean Decrease Gini (MDI) scores are displayed in Table 4.

Table 4. *The 20 most important features of the Random Forest Classifier for each category*

Best Director		Best Picture		Lead acting categories		Supporting acting categories	
Variable name*	MDI	Variable name	MDI	Variable name	MDI	Variable name	MDI
DGA_win	0,329	DGA_win	0,240	GG_drama_lead_win	0,089	GG_supporting_win	0,092
total_oscar_noms	0,092	total_oscar_noms	0,068	SAG_win_2	0,085	rt_critic_score	0,076
gg_win	0,070	rt_audience_score	0,063	rt_audience_score	0,069	imdb_score	0,073
imdb_score	0,064	imdb_score	0,056	rt_critic_score	0,062	rt_audience_score	0,072
rt_critic_score	0,060	PGA_win_2	0,053	imdb_score	0,060	total_oscar_noms	0,071
rt_audience_score	0,057	rt_critic_score	0,052	total_oscar_noms	0,058	BAFTA_nom	0,029
critics_choice_win_2	0,042	GG_drama_win	0,051	BAFTA_win	0,057	SAG_win_2	0,027
BAFTA_win	0,026	critics_choice_win_2	0,035	previous_oscar_noms	0,034	previous_oscar_noms	0,023
BAFTA_nom	0,018	BAFTA_win	0,023	SAG_win_1	0,029	BAFTA_win	0,023
gg_nom	0,015	PGA_win_1	0,022	critics_choice_win_2	0,028	GG_supporting_nom	0,022
critics_choice_win_1	0,013	BAFTA_nom	0,018	Q1_release	0,022	Q4_release	0,022
PG	0,013	SAG_win_2	0,018	best_film_nom	0,022	45-55	0,022
biography	0,013	critics_choice_win_1	0,018	BAFTA_nom	0,017	PG	0,020
director_previous_oscar_nom	0,012	SAG_nom_1	0,015	R	0,016	comedy	0,019
critics_choice_nom_2	0,012	SAG_win_1	0,015	Q4_release	0,016	SAG_win_1	0,019
crime	0,010	Q4_release	0,013	35-45	0,015	best_film_nom	0,018
drama	0,010	SAG_nom_2	0,013	critics_choice_win_1	0,015	R	0,018
Q4_release	0,010	best_dir_nom	0,012	previous_oscar_wins	0,015	Q1_release	0,017
Q1_release	0,009	GG_drama_nom	0,012	GG_comedy_lead_win	0,014	Q3_release	0,016
critics_choice_nom_1	0,009	GG_comedy_win	0,011	SAG_cast_win_1	0,014	critics_choice_win_2	0,016

*A full list of the explanatory variables and their description can be seen in Appendix A

3.3 Support Vector Machines

Firstly I will explain what a support vector classifier is, based on the book of Tibshirani et al. (2013). The support vector classifier is a non-probabilistic binary classifier. It

constructs a hyperplane in high dimensional space. In two dimensions, a hyperplane is simply a line. The equation

$$\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p = 0 \tag{6}$$

defines a p-dimensional hyperplane. In case a point (vector of length p) $X = (X_1, X_2, \ldots, X_p)^T$ satisfies (6), then X lies on the hyperplane. If X does not satisfy the equation, then it lies on either side of the hyperplane. Therefore the hyperplane divides the p-dimensional space into two halves. In the classification setting, our goal is to classify a test observation, a p length vector of the observed features: $x^* = (x_1^*, \ldots, x_p^*)^T$. For this we have our database, a $n \times p$ data matrix of n training observations in p-dimensional space, and these observations fall into two classes: $y_1, \ldots, y_n \in \{-1,1\}$. In my case the two classes 1 and -1 are Oscar winners and non-winners. We make our classifications based on the sign of $f(x^*) = \beta_0 + \beta_1 x_1^* + \ldots + \beta_p x_p^*$. If $f(x^*)$ is positive, then we assign the test observation to class 1, and if $f(x^*)$ is negative, then we assign it to class -1. If $f(x^*)$ is far from zero, then this means that x^* lies far from the hyperplane, and so we can be confident about our class assignment for x^*. To get the best classifier, we want to choose the separating hyperplane that is the farthest away from the training observations. This hyperplane is called the maximal margin classifier, and it is generated by finding the hyperplane with the largest minimal distance to the training observations (the largest margin). A large margin is intended to guarantee that test observations will fall on the right side of the hyperplane, i.e. they will be classified correctly. Observations that are of equal distance to the hyperplane, thus fall on the "edge" of the margins are called support vectors. The maximal margin hyperplane depends directly on these support vectors, hence the name of the model.

In some cases there is no hyperplane that can perfectly separate the two classes. Furthermore, to avoid overfitting, it is worthwhile to misclassify a few training observations in order to do better on the test set. A support vector classifier does exactly this: it uses a soft margin, so it can be violated by some of the training observations. The hyperplane for the support vector classifier is the solution to the optimization problem:

$$\text{maximize } M \tag{7}$$

$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 = 1, \tag{8}$$

$$y_{i(}\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M(1 - \epsilon_i)\,, \tag{9}$$

$$\epsilon_i \geq 0, \sum_{i=1}^{n} \epsilon_i \leq C, \tag{10}$$

Where M is the width of the margin; $\epsilon_i, \ldots, \epsilon_n$ are slack variables, that allow individual observations to be on the wrong side of the hyperplane or the margin; and C is a nonnegative tuning parameter. The ith observation is on the correct side of the margin if $\epsilon_i = 0$, is on the wrong side of the margin if $\epsilon_i > 0$ and is on the wrong side of the hyperplane (misclassified) if $\epsilon_i > 1$. C is the sum of ϵ_i's, thus it determines how much violation should be tolerated. If we want narrow margins, that are rarely violated, a small C is suggested, and vice versa.

To be able to handle non-linear decision boundaries, we have to enlarge the feature space. This is what the support vector machine (SVM) does, which is an extension of the support vector classifier. The solution of the above optimization problem involves only the inner product of the observations, and not the observations themselves. This can be replaced with a generalization of the inner product with the function

$$K(x_i, x_{i'}), \tag{11}$$

called a kernel. A kernel function quantifies the similarity between two observations. For instance if we use

$$K(x_i, x_{i'}) = \sum_{j=1}^{p} x_{ij} x_{i'j}, \tag{12}$$

also called a linear kernel, we get back the original support vector classifier, since the right side of equation (12) is exactly how the inner product is calculated. One can see that the linear kernel measures similarity using simply Pearson correlation. A widely used kernel function is the Gaussian radial basis function:

$$K(x_i, x_{i'}) = exp(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2), \tag{13}$$

where γ is a positive constant. When the support vector classifier is combined with a non-linear kernel, such as the Gaussian kernel, the resulting classifier is a support vector machine. The separating hyperplane in the transformed space has the form

$$f(x) = \beta_0 + \sum_{i \in S} \alpha_i K(x, x_i). \tag{14}$$

Again, for a given test observation $x^* = (x_1^*, \ldots, x_p^*)^T$, the predicted class label will be the sign of $f(x^*)$. If a training observation x_i is far from x^*, then x_i will play little or no role in the predicted class label for x^*. Intuitively, the parameter γ defines how far the influence of a single training observation reaches. A small gamma means that a single training observation x_i will have an influence on deciding the class for x^* even if the distance between them is large. If gamma is large, the influence of x_i is less wide-spread.

In order to find a balanced value for the C and gamma tuning parameters, I implemented *GridSearchCV* in Python, an algorithm that is designed to find the best parameter combination for a given dataset by fitting all the possible combinations and evaluating them. The results of *GridSearchCV* are listed in Table 3.

Table 5. *SVM Parameters by category.*

Parameters	C	γ
Best Picture	100	0,001
Best Director	100	0,001
Lead acting categories	10	0,01
Supporting acting categories	0.1	1

GridSearchCV looked for values in the range of [0.1,1,10,100] for C and [1,0.1,0.01,0.001,0.0001] for γ.

As mentioned before, the support vector machine is a non-probabilistic classifier. However, when predicting Oscar wins I needed probabilities, since my prediction for each year in each category is the film with the highest probability of winning. Getting

probabilities for a support vector machine is implemented in Python' Scikit-learn package: in the binary case probabilities are calibrated using Platt scaling, according to Platt (1999).

Support Vector Machines are generally known as a very powerful tool for classification that can yield accurate predictions in a variety of cases. Another advantage that is beneficial for my classification problem is that SVMs can handle large feature spaces (Sachan, 2015). If the parameters C and γ are appropriately chosen, SVMs can be robust. However, one disadvantage is that SVMs lack the transparency of results (Auria, 2008).

4. Discussion

Predictive modelling of Oscar related data enables considerably accurate predictions of the final outcomes. In this section I will analyze the model's results in scenarios where the model's predictions were far from the reality. I will also evaluate the models' performance for this year's Oscar ceremony. Furthermore, possible improvements to the models will be discussed.

4.1 Surprises and losers

Analyzing misclassifications can offer interesting insights to the classification problem. Following Pardoe's idea (Pardoe, 2005) idea, I looked for the misclassifications where the discrepancy between the actual winner's probability of winning and the predicted winner's probability of winning is largest. For this purpose, I chose to analyze the logit model's estimated probabilities, since the Random Forest classifier performs too well (it only has a single misclassification in the Best Director category) and the Support Vector Machine is a non-probabilistic model. The misclassifications with the three largest discrepancies for each category can be found in Table 6. We can refer to the films on the left side of the table as surprise winners. They were not likely to win according to the model, which means that they probably did not win many of the preceding awards. The nominees on the right side of the table can be referred to as 'losers', since they had a good chance of winning the Oscar, but ended up not doing so.

I tried to look for some online articles to figure out whether the discrepancies that we see in Table 6 are due to the model's errors specifications, or they were considered as surprises by the industry as well. *Business Insider*'s article "*The 20 biggest Oscar upsets of all time*" names two from my list: Anna Paquin's and *Crash*'s win (Guerrasio, 2017). *BBC*'s article with a similar title (*Top 10 biggest Oscar upsets of all time)* mentions four: Frances Ford Coppola not winning Best Director; *Chariots of Fire*, *Braveheart* and *Crash* winning Best Picture. *Crash*'s win seems to have been a huge surprise for the film industry. According to *Indiewire*, "the most notorious Oscar moment of the 21st century is when *Crash* stole Best Picture from *Brokeback Mountain* in 2006" (Sharf, 2018). Russell Crowe not winning in 2002, can probably

be explained with the fact he won Best Actor the year before, for his performance in *Gladiator*. What I draw as conclusion is that at least some of the misclassifications of the models are due to surprises.

Table 6. *Surprises and losers*

Year	Winner	Probability	Predicted	Probability
Best Director				
1969	Carol Reed (Oliver!)	0,21	Anthony Harvey (The Lion in Winter)	0,68
1973	Bob Fosse (Cabaret)	0,33	Francis Ford Coppola (The Godfather)	0,92
2003	Roman Polanski (The Pianist)	0,25	Rob Marshall (Chicago)	0,64
Best Picture				
1982	Chariots of Fire	0,06	Reds	0,88
1996	Braveheart	0,00	Apollo 13	0,99
2006	Crash	0,14	Brokeback Mountain	0,97
Lead acting categories				
1987	Dianne Wiest (Hannah and Her Sisters)	0,15	Maggie Smith (A Room with a View)	0,73
1994	Anna Paquin (The Piano)	0,15	Winona Ryder (The Age of Innocence)	0,53
2007	Alan Arkin (Little Miss Sunshine)	0,09	Eddie Murphy (Dreamgirls)	0,74
Supporting acting categories				
1961	Elizabeth Taylor (BUtterfield 8)	0,01	Shirley MacLaine (The Apartment)	0,62
2002	Denzel Washington (Training Day)	0,01	Russell Crowe (A Beautiful Mind)	0,93
2008	Marion Cotillard (La Vie en Rose)	0,09	Julie Christie (Away from Her)	0,59

4.2 Results for 2018

Let's begin with what needs to be known about this year's Oscars. According to journalists, most Oscars were awarded to the most probable nominees of their category in 2018. Whether they were given to the nominees who deserved them the most, is not my question to answer.

It's safe to say, that there were no big surprises. Nonetheless, prior to the ceremony the fate of the Best Picture award was ambiguous. Ben Zauzmer said, that according to his models, 2018 is the "closest best picture race in at least two decades" (Zauzmer, 2018). *Three Billboards Outside Ebbing, Missouri* won the BAFTA for Best Film, Golden Globe's Best Drama Motion Picture award and the Screen Actors Guild Award for "Outstanding Performance by a Cast in a Motion Picture". Nevertheless, it was not nominated for Best Director, which is not a good prognostic for a film that wants to win the Oscar for Best Picture. *Shape of Water*, on the other hand was nominated for Best Director, won the Directors Guild Award, the Producers Guild Award along with the Critics Choice Award, and it had the most number of total nominations. The statistics slightly favour *Shape of Water*, which in fact turned out to be the winner of the Best Picture Oscar in 2018. Only one of my models predicted this correctly: the random forest classifier which actually managed to yield accurate predictions in all six categories. The estimated probabilities also show how close the race was: the difference between the probability of winning for *Shape of Water* and *Three Billboards Outside Ebbing, Missouri* was only 1.6 percentage points in the random forest model.

The individual predictions for each model can be read from Table 7. Correct predictions are marked in bold in the table. As we can see, predictions were uniform in three categories: Best Director, Best Actor in a Leading Role and Best Actress in a Leading Role. The SVM model's predictions in the supporting categories were incorrect, and due to the models non-transparent, "black-box" nature, I cannot give an explanation as to what might have caused these predictions to go off.

Table 7. *Predictions for each category*

	Actual winner	Logit predictions	Random Forest predictions	SVM predictions
Best Picture	The Shape of Water	Three Billboards Outside Ebbing, Missouri	**The Shape of Water**	Three Billboards Outside Ebbing, Missouri
Best Director	Guillermo del Toro (The Shape of Water)	**Guillermo del Toro** (The Shape of Water)	**Guillermo del Toro** (The Shape of Water)	**Guillermo del Toro** (The Shape of Water)
Best Actor in a Leading Role	Gary Oldman (Darkest Hour)	**Gary Oldman** (Darkest Hour)	**Gary Oldman** (Darkest Hour)	**Gary Oldman** (Darkest Hour)
Best Actress in a Leading Role	Frances McDormand (Three Billboards Outside Ebbing, Missouri)	**Frances McDormand** (Three Billboards Outside Ebbing, Missouri)	**Frances McDormand** (Three Billboards Outside Ebbing, Missouri)	**Frances McDormand** (Three Billboards Outside Ebbing, Missouri)
Best Actor in a Supporting Role	Sam Rockwell (Three Billboards Outside Ebbing, Missouri)	**Sam Rockwell** (Three Billboards Outside Ebbing, Missouri)	**Sam Rockwell** (Three Billboards Outside Ebbing, Missouri)	Christopher Plummer (All the Money in the World)
Best Actress in a Supporting Role	Allison Janney (I, Tonya)	**Allison Janney** (I, Tonya)	**Allison Janney** (I, Tonya)	Mary J. Blige (Mudbound)
Accuracy	–	*84%*	*100%*	*50%*

By analyzing the results of the three most known journalists who use some kind of predictive modelling to forecast the Oscars, I can again conclude that 2018 was not an impossible year to make good predictions. In the six categories that I analyzed, Ben Zauzmer and the analysts at *FiveThirtyEight* made the same (correct) predictions as my random forest model (Zauzmer, 2018; Hickey, 2018). David Rothschild, who built his model based on prediction markets this year, made the same predictions as my logit model - thus also mispredicting the outcome of the Best Picture category (Rothschild, 2018).

4.3 Improvements to the models

The overall accuracy of the models was favourable, especially that of the random forest classifier. Going forward, I would try to build upon that model to make predictions for future Oscar ceremonies. The first obvious point for extension of the model would be to involve more categories. The Best Adapted Screenplay and Best Original Screenplay categories would be reasonable choices to continue, as many already available variables could be used in these categories as well. The other categories are probably more domain specific - a reasonable amount of data collection and data cleaning would be necessary before building any models.

Another interesting question to explore with similar datasets and similar statistical techniques would be predicting Oscar nominations before they are announced. It is presumably a more challenging task, but a successful model would be very useful for the film industry. [4]

Improvements to the existing models might also be made. For instance, it is possible that observations from the earlier years only distort predictions for the present, therefore it would be better to train models only on the more recently nominated films. Finding the best time interval to train the model could be done by first making a train test split so that only very recent movies are in the test set. Second, k different bootstrapped training sets of equal length should be created from the original training set, so that each new training set represents a different time interval (1970 and onwards; 1980 and onwards; etc.). The best time interval could be found by examining which of the k models performs best on the test set using cross validation.

It might also be worthwhile to include new variables for further improvement, especially in the supporting acting categories. It is apparent from Figure 4 and Table 1 that these categories are relatively harder to predict than the other ones I examined. Presently I have no variables that measure 'fame' or 'hype' or star power. These are phenomenons that are difficult to grasp, yet alone measure. Perhaps a way to do it would be to analyze social media mentions, such as tweets or Google searches for film titles and actor names.

---

[4] This is what Balasubramaniam et al. (2014) attempted to do for the Golden Globes, with limited success. Ben Zauzmer also made nomination predictions for the Hollywood Reporter, quite successfully (Zauzmer, 2018b).

New variables could help to be able to add new information into the models that are currently not included. On the contrary, one weakness of my models is the lack of professional feature selection. Stepwise procedures such as backwards elimination aren't built into Scikit-learn, which is why I had to settle for methods such as *SelectKBest* in the logit model for instance. A solution would be to use L1-penalization or to implement an algorithm for backwards elimination in Python myself. Another limitation of my models is that I did not differentiate between male and female actors when considering the variables to be used and when building the models. By handling male and female actors separately, more precise statistical analysis could be performed and perhaps even creating more accurate models would be possible.

Conclusion

It is possible to predict the Academy Awards recipients with considerably high accuracy using statistical learning techniques. I found that for this purpose a random forest classifer is the best performing model, which is an ensemble of decision trees. Its success might be due to the fact that I was essentially trying to model a voting process, and decision trees are known to mirror human behaviour effectively. The random forest model correctly predicted 91.5% of Oscar winners in the six main categories between 1960 and 2018. The overall accuracy of the logit model and the SVM are 70% and 86% respectively. Among the categories that I examined, the award for Best Director proved to be the most predictable and the supporting acting categories the least.

One of the goals of this project was to forecast the Oscars winner for this year. I managed to make good predictions for 2018, similarly to the mathematicians and economists who make forecasts for the film industry. During my analysis, I drew the conclusion, that in certain cases, when the models cannot identify the winner correctly, the recipient of the award is a surprise to the industry too. In 2018, there weren't many surprise winners.

The other goal was to identify the most important factors that can predict the outcome of the Oscars. The most important are the results of preceding award ceremonies, such as the DGA, the BAFTA, the Golden Globe's Drama cluster and the SAG awards. The number of total Oscar nominations and Best Picture or Best Director nominations are also good indicators whether a film would win. Popular and critical acclaim measured by Rotten Tomatoes scores and in some cases the genre classifications of a movie can also be predictive. Several trends can be examined about the Oscars, for example regarding release date, however, these are not as useful for prediction.

REFERENCES

Auria, L. and Moro, R. (2008). Support Vector Machines (SVM) as a Technique for Solvency Analysis. *SSRN Electronic Journal.*

Arrow, K., et al. (2008.)The promise of prediction markets. *Science*, 320(5878), pp. 877-878.

Balasubramaniam, D. K., Chan J., Kulkarni, J. and Zheng D. (2014). Predicting Golden Globe Awards & Christmas Day Movie Gross. *(Final Report: Team 2)*. Stony Brook University. Available at: https://www3.cs.stonybrook.edu/-skiena/591/final_projects/movie_gross/reports/CSE591-Team2-Final_Report.pdf [Accessed 13 Mar. 2018].

Bialik, C. (2013). *And the Oscar-Pool Winners Are...the Stats Dudes.* [online] WSJ. Available at: https://www.wsj.com/articles/SB10001424127887324503204578318682787064790 [Accessed 6 Mar. 2018].

Blauvelt, C. (2016). *The 10 biggest Oscar upsets of all time*. [online] Bbc.com. Available at: http://www.bbc.com/culture/story/20160114-the-10-biggest-oscar-upsets-of-all-time [Accessed 10 Mar. 2018].

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(3), pp.261-277.

Cruwys, S. (2017). *And the Award Goes To*, GitHub repository. Available at: https://github.com/scruwys/and-the-award-goes-to [Accessed 6 Mar. 2018].

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), pp.273-297.

Deuchert, E., Adjamah, K. and Pauly, F. (2005). For Oscar Glory Or Oscar Money?. *Journal of Cultural Economics*, 29(3), pp.159-176.

Guerrasio, J. (2017). *The 20 biggest Oscar upsets of all time — and where the stunning 'Moonlight' win ranks*. [online] Business Insider. Available at: http://www.businessinsider.com/biggest-oscar-upset-winners-2017-2#1-forrest-gump-beats-pulp-fiction-and-the-shawshank-redemption-for-best-picture-1995-20 [Accessed 10 Mar. 2018].

Hickey, W. (2018). *Oscars 2018: Here Are Our Final Predictions*. [online] FiveThirtyEight. Available at: https://fivethirtyeight.com/features/oscars-2018-here-are-our-final-predictions/ [Accessed 13 Mar. 2018].

Hunter, J.D. (2007). Matplotlib: A 2D Graphics Environment, *Computing in Science & Engineering,* 9, pp. 90-95

Internet Movie Database (n.d.). Accessed at: http://www.imdb.com/.

IMDB.com (n.d.). *Names.basics dataset* [tsv]. Accessed at https://datasets.imdbws.com/ [Accessed 11 Jan. 2018].

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013.). *An introduction to statistical learning.*

Kaplan, D. (2006). And the Oscar goes to: A logistic regression model for predicting Academy Award results. *Journal of Applied Economics and Policy,* 25, pp. 23-41.

Maimon, O. and Rokach, L. (2010). *Data Mining and Knowledge Discovery Handbook.* Boston, MA: Springer Science+Business Media, LLC, pp.166-192.

McKinney, W. (2010). Data Structures for Statistical Computing in Python, *Proceedings of the 9th Python in Science Conference*, pp. 51-56.

Nelson, R., Donihue, M., Waldman, D. and Wheaton, C. (2001) "What's an Oscar Worth?" *Economic Inquiry*, 39, pp. 1–16.

Pardoe, I. (2005). Just How Predictable Are The Oscars?. *CHANCE*, 18(4), pp.32-39.

Pardoe, I., & Simonton, D. K. (2008). Applying discrete choice models to predict Academy Award winners. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171, pp. 375–394.

Pedregosa F. et. al. (2011). Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research*, 12, pp. 2825-2830.

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers* 10 (3), pp. 61-74.

Pandya, P. (2015). *The Academy Awards, 1927-2015* [csv]. Available at: https://www.kaggle.com/theacademy/academy-awards [Accessed 13 Mar. 2018].

Pathak D., Rothschild D. and Dudik, M. (2015). A Comparison of Forecasting Methods: Fundamentals, Polling, Prediction Markets, and Experts. *Journal of Prediction Markets*, 9 (2).

Raschka, S. (2016). *Regularization in Logistic Regression: Better Fit and Better Generalization?*. [online] Kdnuggets.com. Available at: https://www.kdnuggets.com/2016/06/regularization-logistic-regression.html [Accessed 17 Mar. 2018].

Rothschild, D. (2018). *Oscars 2018 – PredictWise.* [online] Predictwise.com. Available at: https://predictwise.com/blog/2018/03/oscars-2018/ [Accessed 13 Mar. 2018].

Sachan, L. (2018). *Logistic Regression vs Decision Trees vs SVM: Part II - Edvancer Eduventures.* [online] Edvancer.in. Available at: https://www.edvancer.in/logistic-regression-vs-decision-trees-vs-svm-part2/ [Accessed 10 Mar. 2018].

Sharf, Z. (2018). *'Brokeback Mountain' and 'Crash' Producers Look Back At That Infamous Best Picture Upset: 'This Stuff Is So Ridiculous'.* [online] IndieWire. Available at: http://www.indiewire.com/2018/03/brokeback-mountain-crash-producers-best-picture-upset-oscars-1201934798/ [Accessed 10 Mar. 2018].

Silver, N. (2013). *Oscar Predictions, Election-Style.* [online] FiveThirtyEight. Available at: https://fivethirtyeight.blogs.nytimes.com/2013/02/22/oscar-predictions-election-style/ [Accessed 12 Mar. 2018].

Simonoff, J. and Sparrow, I. (2000). Predicting Movie Grosses: Winners and Losers, Blockbusters and Sleepers. *CHANCE*, 13(3), pp.15-24.

Simonton, D. K. (2005). Cinematic creativity and production budgets: Does money make the movie? *Journal of Creative Behavior*, 39, pp.1–15.

Simonton, D. K. (2007). Is bad art the opposite of good art? Positive versus negative cinematic assessments of 877 feature films. *Empirical Studies of the Arts*, 25, pp. 143–161.

Simonton, D. K. (2009). Cinematic success criteria and their predictors: The art and business of the film industry. *Psychology & Marketing*, 26(5), pp. 400-420.

Surowiecki, J. (2005). *The wisdom of crowds*.

The Economist. (2015). *How Oscar winners are decided*. [online] Available at: https://www.economist.com/blogs/economist-explains/2015/01/economist-explains-14 [Accessed 10 Mar. 2018].

Varma, R. (2017). *Rohan Varma's Answer to What are C and gamma with regards to a support vector machine?* Retrieved from https://www.quora.com/What-are-C-and-gamma-with-regards-to-a-support-vector-machine, [Accessed 10 Mar. 2018].

Walt S., Colbert, S. C. and Varoquaux, G.. (2011). The NumPy Array: A Structure for Efficient Numerical Computation, *Computing in Science & Engineering,* 13, pp. 22-30.

Zauzmer, B. (2018a). *Best Actor - Oscars: The Math Predicts a 'The Shape of Water' Best Picture Win*. [online] The Hollywood Reporter. Available at: https://www.hollywoodreporter.com/lists/oscars-math-predicts-a-shape-water-best-picture-win-1089106/item/best-actor-oscars-2018-ben-zauzmer-math-1089112 [Accessed 10 Mar. 2018].

Zauzmer, B. (2018b). *Best Picture - Oscars: Predicting the Nominees With the Help of a Little Math*. [online] The Hollywood Reporter. Available at: https://www.hollywoodreporter.com/lists/oscars-predicting-nominees-help-a-little-math-1075749/item/best-picture-ben-zauzmer-mathematical-predictions-1075750 [Accessed 13 Mar. 2018].

APPENDIX

Appendix A

Full list of explanatory variables

Variable	Coded name	Description	Categories where used
IMDB Score	`imdb_score`	IMDB Score on a [0,10] scale	All categories
RT Audience Score	`rt_audience_score`	Rotten Tomatoes Audience Score on a [0,100] scale	All categories
RT Critics' Score	`rt_critics_score`	Rotten Tomatoes Critics Score on a [0,100] scale	All categories
DGA win	`DGA_win`	1 if film won DGA award, 0 otherwise	Best Picture; Best Director
DGA nomination	`DGA_nom`	1 if film was nominated for DGA award, 0 otherwise	Best Picture; Best Director
SAG award for cast win 1	`SAG_cast_win_1` or `SAG_win_1`	1 if film was released after 1995 and did not win SAG's Outstanding Performance by a Cast in a Motion Picture award, 0 otherwise	Best Picture; Acting categories
SAG award for cast win 2	`SAG_cast_win_2` or `SAG_win_2`	1 if film won SAG's Outstanding Performance by a Cast in a Motion Picture award, 0 otherwise	Best Picture; Acting categories
SAG award for cast nomination 1	`SAG_cast_nom_1` or `SAG_nom_1`	1 if film was released after 1995 and was not nominated SAG's Outstanding Performance by a Cast in a Motion Picture award, 0 otherwise	Best Picture; Acting categories
SAG award for cast nomination 2	`SAG_cast_nom_2` or `SAG_nom_2`	1 if film was nominated for SAG's Outstanding Performance by a Cast in a Motion Picture award, 0 otherwise	Best Picture; Acting categories
SAG win	`SAG_win`	1 if actor or actress won the SAG award, 0 otherwise	Acting categories
SAG nomination	`SAG_nom`	1 if actor or actress was nominated for the SAG award, 0 otherwise	Acting categories

Variable	Coded name	Description	Categories where used
PGA win 1	PGA_win_1	1 if film was released after 1990 and was not nominated for PGA award, 0 otherwise	Best Picture
PGA win 2	PGA_win_2		Best Picture
PGA nomination 1	PGA_nom_1	1 if film was released after 1990 and was not nominated for PGA award, 0 otherwise	Best Picture
PGA nomination 2	PGA_nom_2	1 if film was released after 1990 and was nominated for PGA award, 0 otherwise	Best Picture
BAFTA win	BAFTA_win	1 if film won the BAFTA in their respective category,, 0 otherwise	All categories
BAFTA nomination	BAFTA_nom	1 if film was nominated the BAFTA in their respective category,, 0 otherwise	All categories
Critics Choice win 1	critics_choice_win_1	1 if film was released after 1995 and did not win the Critics Choice award in their respective category, 0 otherwise	All categories
Critics Choice win 2	critics_choice_win_2	1 if film was released after 1995 and won the Critics Choice award in their respective category, 0 otherwise	All categories
Critics Choice nomination 1	critics_choice_nom_1	1 if film was released after 1995 and was not nominated for Critics Choice award, 0 otherwise	All categories
Critics Choice nomination 2	critics_choice_nom_2	1 if film was released after 1995 and was nominated for Critics Choice award in their respective category,, 0 otherwise	All categories
Golden Globe: Best Picture - Drama win	GG_drama_win	1 if film won the Golden Globe for Best Picture (Drama), 0 otherwise	Best Picture
Golden Globe: Best Picture - Drama nomination	GG_drama_nom	1 if film was nominated for Golden Globe for Best Picture (Drama), 0 otherwise	Best Picture
Golden Globe: Best Picture - Musical or Comedy win	GG_comedy_win	1 if film won the Golden Globe for Best Picture (Musical or Comedy), 0 otherwise	Best Picture
Golden Globe: Best Picture - Musical or Comedy nomination	GG_comedy_nom	1 if film was nominated for Golden Globe for Best Picture (Musical or Comedy), 0 otherwise	Best Picture

Variable	Coded name	Description	Categories where used
Golden Globe: Best Actor or Actress in leading role - Drama win	`GG_drama_lead_win`	1 if actor or actress won the Golden Globe for Best Lead Actor or Actress (Drama), 0 otherwise	Lead acting categories
Golden Globe: Best Actor or Actress in leading role - Drama nomination	`GG_drama_lead_nom`	1 if actor or actress was nominated for Golden Globe for Best Lead Actor or Actress (Drama), 0 otherwise	Lead acting categories
Golden Globe: Best Actor or Actress in leading role - Musical or Comedy win	`GG_comedy_lead_win`	1 if actor or actress won the Golden Globe for Best Lead Actor or Actress (Musical or Comedy), 0 otherwise	Lead acting categories
Golden Globe: Best Actor or Actress in leading role - Musical or Comedy nomination	`GG_comedy_lead_nom`	1 if actor or actress was nominated for Golden Globe for Best Lead Actor or Actress (Musical or Comedy), 0 otherwise	Lead acting categories
Golden Globe win	`gg_win` or `GG_supporting_win`	1 if actor or director won the Golden Globe, 0 otherwise	Best Director and acting categories
Golden Globe nomination	`gg_nom` or `GG_supporting_nom`	1 if actor or director was nominated the Golden Globe, 0 otherwise	Best Director and acting categories
Total Oscar nominations	`total_oscar_noms`	Number of total Oscar nomination for the given film	All categories
Best Director nomination	`best_dir_nom`	1 if film was nominated for Best Director Oscar, 0 otherwise	Best Picture
Best Picture nomination	`best_film_nom`	1 if film was nominated for Best Picture Oscar 0 otherwise	Best Director and acting categories
Previous nominee	`previous_nominee`	1 if actor or actress was nominated for the Oscar before, 0 otherwise	Acting categories
Previous winner	`previous_winner`	1 if actor or actress had won the Oscar before, 0 otherwise	Acting categories

Variable	Coded name	Description	Categories where used
Previous nominations	`previous_oscar_noms`	Number of total previous Oscar nominations for an actor or director	Best Director and acting categories
Previous wins	`previous_oscar_wins`	Number of total previous Oscar wins for an actor or director	Best Director and acting categories
Q1 release	`Q1`	1 if film was released between January and March, 0 otherwise	All categories
Q2 release	`Q2`	1 if film was released between April and June, 0 otherwise	All categories
Q3 release	`Q3`	1 if film was released between July and September, 0 otherwise	All categories
Q4 release	`Q4`	1 if film was released between October and December, 0 otherwise	All categories
G	`G`	1 if film's MPAA rating is G (General audiences) , 0 otherwise	All categories
PG	`PG`	1 if film's MPAA rating is PG (Parental Guidence), 0 otherwise	All categories
PG-13	`PG-13`	1 if film's MPAA rating is PG-13 (Parents Strongly Cautioned), 0 otherwise	All categories
R	`R`	1 if film's MPAA rating is R (Restricted), 0 otherwise	All categories
Action	`action`	1 if classified as Action by IMDB, 0 otherwise	All categories
Adventure	`adventure`	1 if classified as Adventure by IMDB, 0 otherwise	All categories
Biography	`biography`	1 if classified as Biography by IMDB, 0 otherwise	All categories
Comedy	`comedy`	1 if classified as Comedy by IMDB, 0 otherwise	All categories
Crime	`crime`	1 if classified as Crime by IMDB, 0 otherwise	All categories
Drama	`drama`	1 if classified as Drama by IMDB, 0 otherwise	All categories

Variable	Coded name	Description	Categories where used
Horror	horror	1 if classified as Horror by IMDB, 0 otherwise	All categories
Family	family	1 if classified as Family by IMDB, 0 otherwise	All categories
Fantasy	fantasy	1 if classified as Fantasy by IMDB, 0 otherwise	All categories
Sci-fi	sci-fi	1 if classified as Sci-fi by IMDB, 0 otherwise	All categories
Sport	sport	1 if classified as Sport by IMDB, 0 otherwise	All categories
Mystery	mystery	1 if classified as Mystery by IMDB, 0 otherwise	All categories
Music	music	1 if classified as Music by IMDB, 0 otherwise	All categories
Romance	romance	1 if classified as Romance by IMDB, 0 otherwise	All categories
History	history	1 if classified as History by IMDB, 0 otherwise	All categories
Thriller	thriller	1 if classified as Thriller by IMDB, 0 otherwise	All categories
Western	western	1 if classified as Western by IMDB, 0 otherwise	All categories
War	war	1 if classified as War by IMDB, 0 otherwise	All categories
<25	<25	1 if actor or actress is younger than 25 years old, 0 otherwise	Acting categories
25-35	25-35	1 if actor or actress is minimum 25 and maximum 34 years old than, 0 otherwise	Acting categories
35-45	35-45	1 if actor or actress is minimum 35 and maximum 44 years old than, 0 otherwise	Acting categories
45-55	45-55	1 if actor or actress is minimum 45 and maximum 54 years old than, 0 otherwise	Acting categories
55-65	55-65	1 if actor or actress is minimum 55 and maximum 64 years old than, 0 otherwise	Acting categories

Variable	Coded name	Description	Categories where used
65-75	65–75	1 if actor or actress is minimum 65 and maximum 74 years old than, 0 otherwise	Acting categories
>75	>75	1 if actor or actress is minimum 75 years old, 0 otherwise	Acting categories